UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
ESPECIALIZAÇÃO EM BANCO DE DADOS

LUIZ FERNANDO TREVISAN

REVISÃO DE MÉTODOS PARA ANÁLISE DE AGRUPAMENTO DE DADOS EM *DATA MINING*

MONOGRAFIA DE ESPECIALIZAÇÃO

PATO BRANCO
2017

LUIZ FERNANDO TREVISAN

REVISÃO DE MÉTODOS PARA ANÁLISE DE AGRUPAMENTO DE DADOS EM *DATA MINING*

Trabalho de Conclusão de Curso, apresentado ao II Curso de Especialização em Banco de Dados, da Universidade Tecnológica Federal do Paraná, campus Pato Branco, como requisito parcial para obtenção do título de Especialista.

Orientador: Prof. Dalcimar Casanova

PATO BRANCO

2017

TERMO DE APROVAÇÃO

REVISÃO DE MÉTODOS PARA ANÁLISE DE AGRUPAMENTO DE DADOS EM DATA MINING

por

LUIZ FERNANDO TREVISAN

Este Trabalho de Conclusão de Curso foi apresentado em 23 fevereiro de 2017 como requisito parcial para a obtenção do título de Especialista em Banco de Dados. O(a) candidato(a) foi arguido(a) pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

_____
Dalcimar Casanova
Prof.(a) Orientador(a)

_____
Marco Antonio de Castro Barbosa
Membro titular

_____
Viviane Dal Molin de Souza
Membro titular

"O Termo de Aprovação assinado encontra-se na Coordenação do Curso"

RESUMO

TREVISAN, Luiz Fernando. Revisão de métodos para análise de agrupamento de dados em *data mining*. 2017. 28f. Monografia (II Curso de Especialização em Banco de Dados) - Universidade Tecnológica Federal do Paraná. Pato Branco, 2017.

Os componentes centrais da tecnologia de mineração de dados estão em desenvolvimento há décadas. Hoje, a maturidade dessas técnicas, aliada aos motores de banco de dados de alto desempenho e aos amplos esforços de integração de dados, tornam essas tecnologias práticas para os ambientes atuais. A análise de agrupamento tem como objetivo separar objetos em grupos, agrupando-os de acordo com as suas características em comum com um critério pré-determinado, identificado padrões compreensíveis. Para realizar esta classificação as diversas técnicas de mineração de dados utilizam funções matemáticas complexas. Nesse contexto, até mesmo uma abstração mais fácil das formulas para agrupamento de dados não é simples de ser entendida, principalmente para quem não é da área ou não tem conhecimento de conceitos matemáticos. O objetivo deste trabalho é esclarecer as fórmulas de alguns métodos de agrupamento de dados, explicando-os de forma pratica e objetiva, com exemplos, de como eles funcionam. Para isso foram escolhidos 3 algoritmos do mesmo gênero, *k-means, k-medians* e *k-medoids,* para serem detalhados utilizando o mesmo conjunto de dados.

Palavras-chave: Agrupamento de Dados. Mineração de Dados.

ABSTRACT

TREVISAN, Luiz Fernando. Review of methods for data clustering analysis in *data mining*. 2017. 28f. Monography (II Specialization Course in Database) - Federal University of Technology - Parana. Pato Branco, 2017.

The core components of data mining technology have been in development for decades. Today, the maturity of these techniques, coupled with high-performance database engines and extensive data integration efforts, make these technologies practical for today's environments. The cluster analysis aims to separate objects into groups, grouping them according to their characteristics in common with a predetermined criterion, identifying comprehensible patterns. To perform this classification the various data mining techniques use complex mathematical functions. In this context, even an easier abstraction of the formulas for grouping data is not simple to understand, especially for those who are not from the area or have no knowledge of mathematical concepts. The purpose of this work is to clarify the formulas of some methods of grouping data, explaining them in a practical and objective way, with examples, of how they work. For this, 3 algorithms of the same genre, k-means, k-medians and k-medoids were chosen to be detailed using the same set of data.

Palavras-chave: Data Clustering. Data Mining.

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1. INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

Com o avanço tecnológico trazendo uma maior coleta e armazenamento em grandes bases de dados, que integram informações operacionais de clientes, fornecedores e de mercado, resultou em uma explosão de informações. A concorrência exige uma análise oportuna e sofisticada sobre uma visão integrada dos dados. No entanto, há uma diferença crescente entre sistemas de armazenamento e a capacidade dos usuários de analisar e agir efetivamente sobre as informações que contêm, um novo salto tecnológico foi necessário para estruturar e priorizar a informação para problemas específicos do usuário final, utilizando da mineração de dados para dar esse salto.

Todo o processo de mineração de dados não pode ser concluído em uma única etapa. Em outras palavras, não é possível obter as informações necessárias dos grandes volumes de dados tão facilmente. É um processo muito complexo que envolve uma série de etapas: limpeza de dados, integração de dados, seleção de dados, transformação de dados, mineração de dados, avaliação de padrões e representação de conhecimento que devem ser concluídos nessa ordem.

Em aplicações de mineração de dados e descoberta de conhecimento em bases de dados, a análise de agrupamento, se encaixa como uma parte da solução para um problema maior, que envolve outras etapas e técnicas (JAIN, TOPCHY, LAW, BUHMANN, 2004). Porém para a análise de agrupamento não existem métodos gerais que sejam adaptáveis a vários tipos de dados e geometrias dos conjuntos de dados. Essas dificuldades tornam a utilização desses métodos bastante complexas e propensa a erros, especialmente de pessoas que não estão familiarizadas com essa área da ciência.

Assim, optou-se nesse trabalho detalhar de forma didática alguns métodos de agrupamento, abstraindo conceitos avançados de matemática e estatística. Isso é importante porque nem todos que querem empregar alguma técnica de agrupamento tem o conhecimento suficiente para interpretar esse tipo de conteúdo

1.2 OBJETIVOS

1.2.1 Objetivo Geral

- Abstrair o conceito matemático dos métodos para agrupamento de dados *k-means* e suas duas principais variações, *k-medoids* e *k-medians*;

1.2.2 Objetivos Específicos

- Escolher um conjunto de dados para fazer os testes;
- Pré-processar o conjunto de dados;
- Implementar os métodos *k-means*, *k-medoids* e *k-medians*;
- Analisar os resultados obtidos nos 3 métodos;
- Apresentar de forma objetiva, com exemplos numéricos o funcionamento dos 3 métodos, suas formulas, tipos de dados permitidos, e as etapas para gerar o agrupamento.

1.3 JUSTIFICATIVA

As técnicas de agrupamento de dados estão implementadas e prontas para o uso em vários pacotes de softwares e linguagens de programação, além de que muitas outras áreas do conhecimento estão interessadas na obtenção dos resultados obtidos pelo agrupamento, e isso vem gerando uma grande demanda, que aliada com a falta de profissionais, faz com que pessoas sem conhecimento adequado tentem faze-la, com isso gerando diversos erros, erros de design de experimentos e principalmente de interpretação dos resultados.

O detalhamento do funcionamento de métodos para agrupamento de dados se justifica pelas diversas áreas que utilizam dessa técnica de sem conhecer as particularidades dos métodos utilizados para classificar os dados coletados, apenas olham o resultado, sem poder interpretar se o método e as métricas aplicadas estão corretas, pois não é fornecido um conjunto claro de diretrizes para indicar como a análise deve ser realizada.

Assim, cabe ao pesquisador determinar qual o método de análise que fornecerá as conclusões mais válidas. Por isso a validade dos resultados obtidos é muitas vezes questionada, uma vez que existe uma série de fatores que influenciam no resultado do agrupamento: os métodos atuais não abordam todos os tipos de dados adequadamente e simultaneamente, lidar com um grande número de dimensões e uma grande quantidade de elementos poder ser problemático devido à complexidade do tempo, a eficácia do método depende diretamente da medida de similaridade utilizada, não existe uma medida de distância óbvia , é necessário defini-la, o que nem sempre é fácil, especialmente em espaços multidimensionais e por fim, o resultado do agrupamento pode ser interpretado de diferentes maneiras (MAHDAVI M., ABOLHASSANI, 2009).

Além de que, para realizar esta classificação as diversas técnicas de mineração de dados utilizam funções matemáticas complexas. Até mesmo a mais fácil das formulas para agrupamento

de dados não é simples de ser entendida, principalmente para quem não é da área ou não tem conhecimento de conceitos matemáticos.

1.4 ESTRUTURA DO TRABALHO

Este texto está organizado em capítulos, dos quais este é o primeiro e apresenta a ideia e o contexto do trabalho, incluindo os objetivos e a justificativa.

No Capítulo 2 está o referencial teórico utilizado como base para a execução deste trabalho.

No Capítulo 3 estão os materiais e o método utilizado e os procedimentos realizados.

O Capítulo 4 contém o detalhamento dos métodos estudados, com exemplos numéricos da implementação de cada um. Os métodos são exemplificados pela apresentação de imagens.

No Capítulo 5 está a conclusão com as considerações finais.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta as teorias que embasam esta pesquisa. A primeira parte consiste de uma revisão da literatura de mineração de dados. Em seguida será abordada mais especificamente a perspectiva relacionada a agrupamento de dados. Seu objetivo é identificar e apresentar os conhecimentos, as definições e as teorias já formuladas.

2.1 MINERAÇÃO DE DADOS

A mineração de dados é o processo de transformar dados brutos em informações úteis. Quaisquer números, textos, fatos, páginas da web ou documentos que podem ser processados por um computador são considerados dados. O processo de mineração permite a análise de dados com muitas dimensões e ângulos diferentes, consegue categorizá-los e resumir as relações entre os elementos, para resumi-los em informação útil, informação que pode ser usada para aumentar a receita, redução de custos, encontrar grupos de clientes com comportamento semelhante, classificar plantas e animais, etc. Tecnicamente, a mineração de dados é o processo de encontrar correlações ou padrões, entre dezenas de campos em grandes bases de dados (BERRY, LINOFF, 1999).

As etapas para mineração de dados podem ser classificadas em dois grupos:

A) Preparação de dados ou pré-processamento de dados:

1. Limpeza de dados: É o processo onde os dados são limpos. Os dados disponíveis em fontes de dados podem estar com valores de atributos de interesse faltantes, ou, às vezes, os dados podem conter erros, um exemplo é um atributo de idade com valor 200, é óbvio que o valor de idade está errado neste caso. Se os dados não estiverem limpos, os resultados da mineração de dados não serão confiáveis nem precisos.
2. Integração de dados: É o processo onde os dados de diferentes fontes de dados são integrados. Os dados podem ser armazenados em bancos de dados, arquivos de texto, planilhas, documentos, cubos de dados, Internet e assim por diante. Um dos problemas enfrentados é a redundância de dados, os mesmos dados podem estar disponíveis em tabelas diferentes no mesmo banco de dados ou mesmo em diferentes fontes de dados.
3. Seleção de dados: A seleção de dados é o processo onde os dados relevantes para a análise são recuperados da base de dados, o processo de mineração de dados requer grandes volumes de dados históricos para

análise, a partir dos dados disponíveis, os dados de interesse precisam ser selecionados e armazenados.

4. Transformação de dados: É o processo de transformar e consolidar os dados em diferentes formas que são adequados para a mineração. A transformação de dados envolve normalmente normalização, agregação, generalização. Aqui os dados se tornam mais adequados para a mineração de dados.

Após a integração dos dados, os dados disponíveis estão prontos para a mineração de dados.

B) Mineração de dados:

1. Mineração de dados: É o processo central onde uma série de métodos complexos e inteligentes são aplicados para extrair padrões. O processo de mineração de dados inclui uma série de tarefas, tais como associação, classificação, previsão, agrupamento, análise de séries temporais e assim por diante.
2. Avaliação de padrões: Identifica os padrões verdadeiramente interessantes que representam o conhecimento baseado em diferentes tipos de medidas de interesse. Um padrão é considerado interessante se for potencialmente útil e facilmente compreensível por humanos.
3. Representação de conhecimento: As informações extraídas dos dados precisam ser apresentadas ao usuário de forma atraente. Diferentes técnicas de visualização e representação de conhecimento são aplicadas para fornecer a saída do data mining para os usuários.

A **Figura 1** demonstra as etapas da mineração de dados, separadas nos dois grupos.

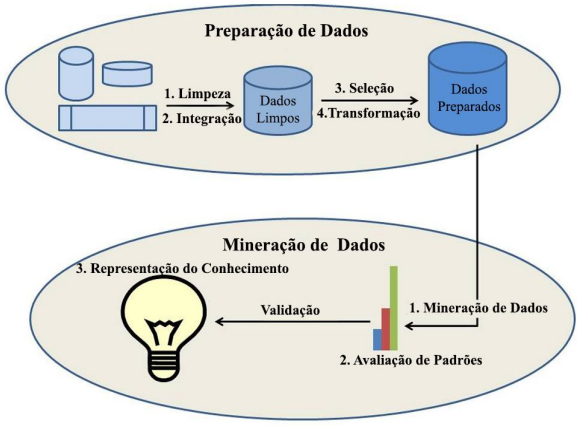

Figura 1 Etapas do processo de mineração de dados

2.2 AGRUPAMENTO DE DADOS

O processo de agrupamento pode ser definido como a identificação de um conjunto de categorias (usualmente chamadas de grupos ou *clusters*) que descrevem um conjunto de elementos (FAYYAD, DJORGOVSKI, WEIR, 1996). Ainda, Conforme MACQUEEN, 1967, o objetivo desta divisão em grupos é fazer com que elementos do mesmo *cluster* sejam mais homogêneos e, ao mesmo tempo, fazer com que os elementos de *clusters* distintos se tornem mais heterogêneos.

Existem inúmeros algoritmos baseados em medidas de similaridade para a formação dos agrupamentos, os métodos de agrupamento exigem de seus usuários a tomada de diversas decisões, fazendo surgir à necessidade do conhecimento das propriedades de cada um deles. Conhecendo-se os diferentes algoritmos, é possível determinar qual deles é a melhor escolha para atingir os objetivos estabelecidos, fazendo assim com que o tempo e o custo de recuperação diminuam (KAUFMAN, ROUSSEEUW, 1987).

2.2.1 K-MEANS

O método *k-means* é um dos mais simples dentre os algoritmos de agrupamento. Ele não é muito flexível, e com isso tem seu uso limitado. Ele segue uma maneira simples e fácil de classificar um conjunto de dados em *k clusters* (MACQUEEN J. B., 1967). O algoritimo tem sua execução alternando entre dois passos principais:

1) Passo de Atribuição: Atribuir cada elemento do conjunto de dados ao grupo, cuja a distância seja a "mais próxima", sendo cada elemento é atribuído a um único grupo mesmo se pudesse ser atribuído a dois ou mais deles.
2) Passo de atualização: Calcula-se os novos centros de grupo. A forma de obtenção do novo centro é diferente para cada variação dos algoritmos estudados.

Esse método só pode ser utilizado com conjuntos de dados que contenham atributos de tipos compatíveis com a média, um atributo categórico, como por exemplo a situação do clima, nublado, ensolarado ou chuvoso, não poderia ser agrupada pelo método k-means.

2.2.2 K-MEDIANS

O algoritmo k-medians é uma variação do k-means, que em vez de utilizar o cálculo da média dos elementos para determinar o seu centro, utiliza o cálculo da mediana. Como a média é uma medida vulnerável que pode puxar o resultado para longe da maioria do conjunto de dados, a

mediana, por outro lado, é uma estatística incrivelmente menos susceptível a valores excepcionalmente altos ou baixos.

2.2.3 K-MEDOIDS

Em contraste com o algoritmo k-means, o k-medoids tem alguns pontos que se destacam. O seu ponto positivo, é sua robustez ao ruído, pois o centro do grupo é definido como um elemento do próprio grupo, isso significa um ponto mais centralmente localizado no grupo. Já seu ponto negativo é sua necessidade computacional mais elevada, normalmente, o k-medoids demora mais tempo para ser executado, o que pode ser um problema ao se trabalhar com enormes quantidades de dados.

2.3 DISSIMILARIDADE ENTRE ELEMENTOS

Para determinar quanto um elemento é parecido com outro, esses métodos utilizam medidas de distância, é muito importante conhecer o tipo de medida utilizado pelo método aplicado no conjunto de dados, pois a escolha de diferentes tipos de media influenciará diretamente no resultado final do agrupamento.

É necessário escolher a medida analisando os tipos de dados que serão minerados, qual tipo de medida utilizar para elementos que contenham características em escalas de valor diferentes? Qual tipo de medida utilizar para características categóricas ou medidas não numéricas? Já que os métodos estudados podem ser executados utilizando vários tipos de medidas de dissimilaridade.

Foi escolhida a distância euclidiana (ou distância métrica) pois é uma noção natural de distância (é a menor distância entre dois pontos), e também por todas as características dos elementos estarem na mesma escala.

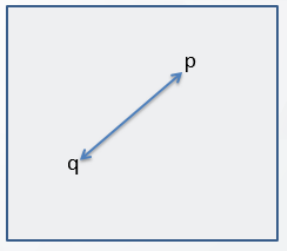

Figura 2 Distancia euclidiana.

Outras formas de medida também podem ser utilizadas, como por exemplo:

a) Distância Levenshtein ou distância de edição, essa medida de distância é utilizada para calcular a distância entre duas palavras ou *strings.* Onde é necessário medir o esforço para transformar uma palavra em outra, aonde cada ação feita para modificação tem o peso 1.

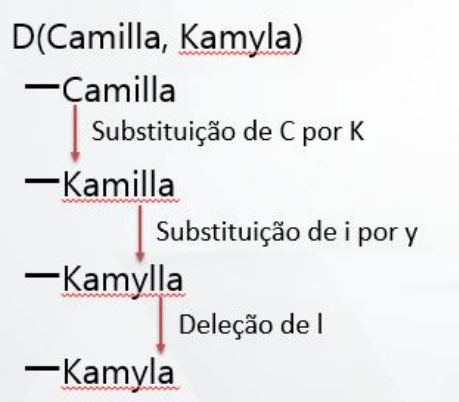

Figura 3 Exemplo da distância de edição entre as palavras Camilla e Kamyla

Assim a distância entre as palavras Camilla e Kamyla é 3.

b) Distância de Mahalanobis, essa medida é utilizada para determinar a distância entre elementos onde as suas características possuam em diferentes escalas, como por exemplo um conjunto de dados cartográficos com valores em km e cm, assim sem a necessidade de normalização dos dados.

3 AQUISIÇÃO DE DADOS

O conjunto de dados Iris foi criado por Ronald Fisher, é um conjunto de dados com 50 amostras de cada uma das 3 espécies, com a medida de 4 características para cada amostra (altura e largura da sépala, altura e largura de pétala) e foi extraído do repositório de aprendizado de maquinas, UCI *Machine Learning Repository* (LICHMAN, M. (2013))

A **Tabela 1** contém uma linha de cada uma das 3 espécies existentes no conjunto de dados.

Fisher's Iris dataset				
Sépala altura	Sépala largura	Pétala altura	Pétala largura	Espécie
5.1	3.5	1.4	0.2	I. setosa
4.9	2.4	3.3	1.0	I. versicolor
6.0	2.2	5.0	1.5	I. virginica

Tabela 1 Exemplo de valores no conjunto de dados Iris

A **Figura 4** apresenta uma foto de cada uma das 3 espécies de flores Iris existentes no conjunto de dados.

Figura 4 As três espécies presentes no conjunto de dados

3.1 PRÉ-PROCESSAMENTO

Para que fosse possível uma melhor visualização dos dados foi necessário um pré-processamento no conjunto de dados iris original. Foi multiplicado os valores de altura e largura da pétala e sépala, assim é possível visualizar os elementos em duas dimensões, fazendo com que o objetivo de explicar os métodos de forma prática e objetiva seja alcançado.

A **Tabela 2** contém uma linha de cada uma das 3 espécies existentes no conjunto de dados Iris após o pré-processamento.

Iris dataset modificado	
Sépala altura * largura	Pétala altura * largura
13,20	7.50
11.76	3.30
18.90	0.28

Tabela 2 Linha do conjunto de dados após o pré-processamento

A **Figura 5** apresenta uma exibição de todos os dados em um gráfico 2D após a realização dos procedimentos de pré-processamento.

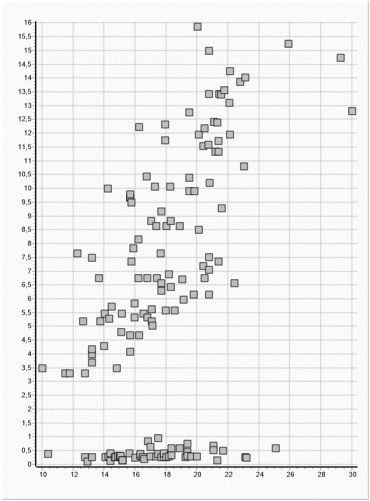

Figura 5 elementos do conjunto de dados iris modificado exibido em 2D

4 AGRUPAMENTO DE DADOS

O Embarcadero RAD Studio juntamente com a biblioteca de componentes de gráficos do TeeChart, foi utilizado para implementação, na linguagem Delphi, dos 3 métodos de agrupamento estudados neste trabalho.

Cada um dos 4 passos necessários para geração do agrupamento será descrito nas subseções a seguir. A última subseção irá apresentar os resultados obtidos com este detalhamento.

4.1 ESCOLHA DO NÚMERO DE GRUPOS EM QUE O CONJUNTO SERÁ DIVIDIDO

O primeiro passo para o agrupamento é escolher a quantidade de grupos que em que o conjunto de dados será dividido. Para os métodos estudados, esse valor de *k* (ou grupos em que se deseja dividir o conjunto) embora seja muito importante, é fornecida pelo usuário (KAUFMAN, ROUSSEEUW, 1990). Algumas das formas de escolher o valor de grupos para divisão do conjunto segundo HAMERLY, ELKAN, 2003:

a) Com base em suposições feitas sobre conjunto de dados;
b) Experiência anterior com um conjunto semelhante;
c) Conhecimento prévio sobre o conteúdo do conjunto;
d) Comparando os resultados obtidos com diferentes números de *k*;

Dessa forma, estes algoritmos se concentram em obter *k* grupos de elementos semelhantes de acordo com um critério pré-estabelecido, que pode ou não produzir o agrupamento desejado.

Quando o número de grupos escolhido é maior que o número de grupos existentes, pode-se enfrentar alguns problemas, como por exemplo:

a) O quarto grupo se torna pequeno ou vazio.
b) O quarto grupo quando posicionado ao centro do conjunto de dados, pode tirar elementos de todos os outros grupos
c) Um dos grupos subdivide-se em dois.

4.2 DEFINIR CENTROS INICIAIS

Após definir a quantidade de grupos em que o conjunto de dados será dividido, a próxima etapa é definir os pontos iniciais para cada um dos *k* centros de grupo.

A **Figura 6** apresenta os 3 pontos iniciais que foram gerados aleatoriamente para o detalhamento dos métodos.

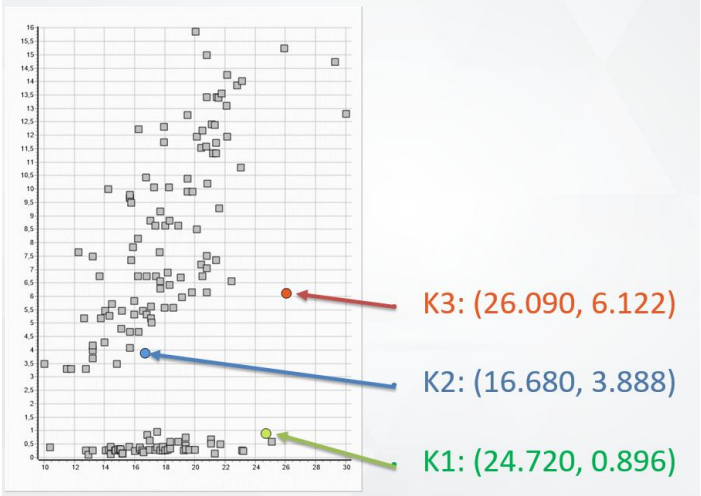

Figura 6 Pontos de inicialização dos centros

O resultado da execução é muito sensível aos pontos de inicialização de seus centros de grupos *k*, cada centro tem a tendência de permanecer mais ou menos no mesmo grupo em que é colocado pela primeira vez (BUBECK, MEILA, VON LUXEMBOURG, 2012).

Sendo assim a **Figura 7** ilustra como a diferente inicialização para o mesmo conjunto de dados pode retornar resultados completamente diferentes.

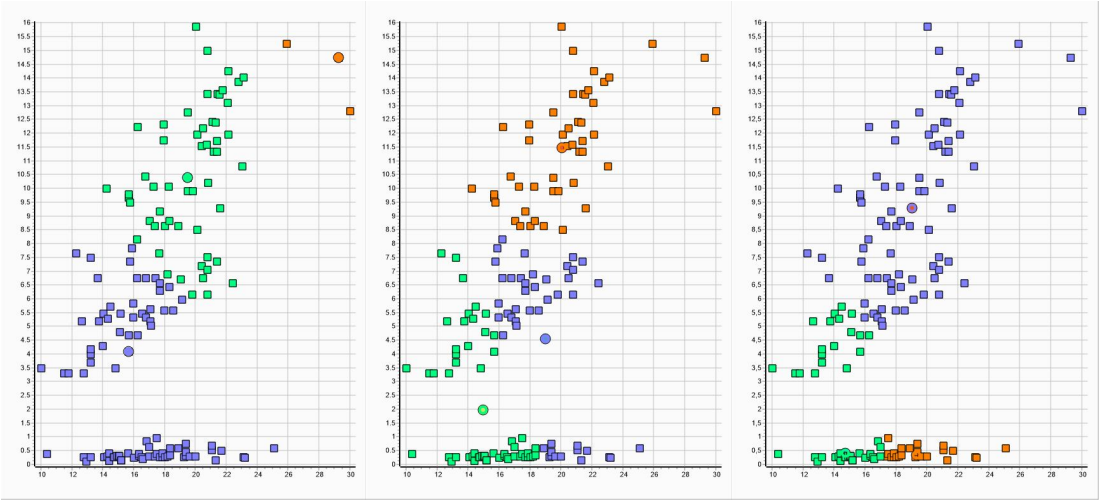

Figura 7 Ilustrações do resultado gerado por diferentes inicializações

4.3 ATRIBUIÇÃO DE CADA ELEMENTO AO GRUPO “MAIS PROXIMO”

Nesse passo é atribuído cada elemento do conjunto de dados ao grupo, cuja a distância euclidiana seja a menor. Onde cada elemento é atribuído exatamente a um único grupo mesmo se pudesse ser atribuído a dois ou mais deles. A distância euclidiana é utilizada como medida de dissimilaridade neste estudo, embora outras medidas possam ser adotadas.

Para exemplificar o cálculo da distância, foi escolhido um ponto definido como p1 e com valores (16.320 e 0.320). Aplicando a formula da distância euclidiana do ponto p1 a cada centro de grupo é obtido os seguintes valores:

A **Figura 8** apresenta a localização do ponto p1 dentre todos os elementos

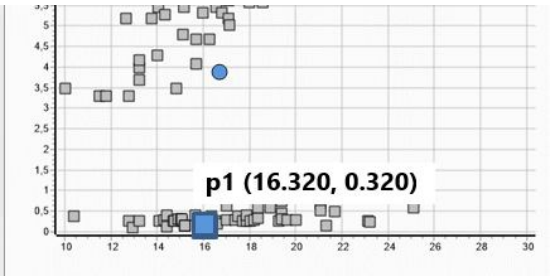

Figura 8 Localização do ponto p1 no gráfico

a) Distância de p1 até k1

$$\sqrt{(16.680 - 16.32)^2 + (3.888 - 0.32)^2} = 3.5861$$

b) Distância de p1 até k2

$$\sqrt{(24.720 - 16.32)^2 + (0.8959 - 0.32)^2} = 8.4197$$

c) Distância de p1 até k3

$$\sqrt{(26.09 - 16.32)^2 + (6.122 - 0.32)^2} = 11.3629$$

Com esses resultados, o elemento p1 é atribuído ao grupo denominado k2, pois sua distância até ele é de 3.5861.

A **Figura 9** exemplifica o cálculo da distância do ponto p1 para cada um dos centros de grupos.

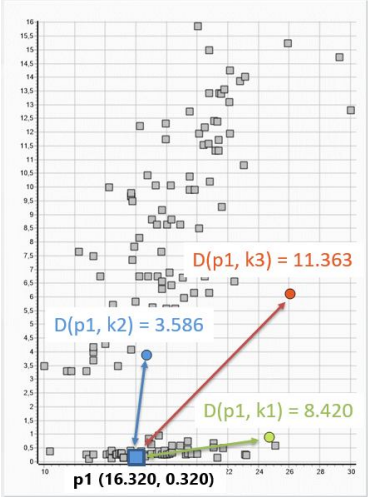

Figura 9 cálculo da distância de um ponto para os centros de grupos

A **Figura 10** ilustra como o conjunto de dados está agrupado após o cálculo e atribuição de todos os elementos do conjunto de dados.

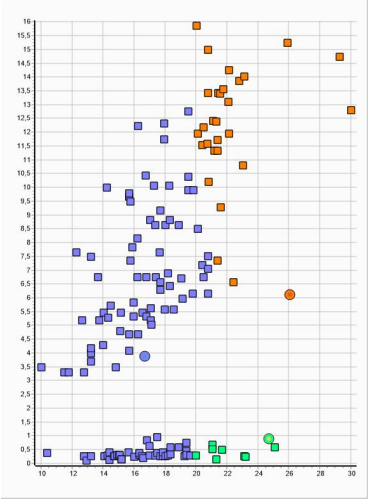

Figura 10 Agrupamento obtido após a atribuição de todos os elementos ao grupo mais próximo

4.4 ATUALIZAÇÃO DOS CENTROS

Nessa etapa será calculado o novo centro de cada um dos grupos, baseado em todos os elementos pertencentes ao grupo. Aqui começam as diferenças entre os métodos. As subseções a seguir irão tratar as particularidades de cada um dos métodos para a atualização dos centros, lembrando que é necessário o cálculo do novo centro individualmente para cada um dos grupos. Será feito o detalhamento do cálculo de atualização utilizando um dos grupos.

A **Tabela 3**, contém todos os 8 elementos pertencentes ao grupo k1, que será o grupo utilizado para exemplificação dessa etapa.

Grupo k1

Sépala	Pétala
21.060	0.680
25.080	0.600
21.060	0.520
21.660	0.510
21.320	0.150
23.100	0.280
23.200	0.240
19.980	0.300

Tabela 3 Grupo k1 formado após primeira execução

4.4.1 K-MEANS

O método *k-means* utiliza a média simples de todos os elementos do cluster, que é divisão da soma de todos os números pela quantidade. Por isso o k-means só pode ser executado com valores de atributos que sejam compatíveis com a média.

Calculo realizado pelo método k-means:

a) Novo valor de *xk* (atributo referente a Sépala)
   a. Soma de todos os valores do atributo

      xk = (21.06 + 19.98 + 23.20 + 25.08 + 21.06 + 21.66 + 21.32 + 23.10)
   b. Divide-se pela quantidade de elementos

      xk = (176.46 / 8)
   c. Novo valor de *xk*

      xk = 22.060

b) Novo valor de *yk* (atributo referente a Sépala)
   a. Soma de todos os valores do atributo

      yk = (0.68 + 0.3 + 0.24 + 0.6 + 0.52 + 0.51 + 0.15 + 0.28)
   b. Divide-se pela quantidade de elementos

      yk = (3.28 / 8)
   c. Novo valor de *xk*

      yk = 0.41

Assim no novo centro do grupo k1 é representado pelos valore (22.060, 0.41)

4.4.2 K-MEDIANS

O método *k-medians* utiliza a mediana de todos os elementos do cluster, a mediana é definida pelo valor que separa a metade maior da metade menor do conjunto. Para fazer o cálculo da mediana, primeiro ordena-se os elementos, depois localiza o elemento que divide o conjunto. Quando o conjunto tem o número de elementos par, a mediana será a média dos dois elementos que separam a metade maior da metade menor. Sendo assim a mediana pode ser utilizada para conjuntos de dados não numéricos.

A **Figura 11** representa o cálculo realizado pelo método *k-medians* para atualização do centro de grupo k1.

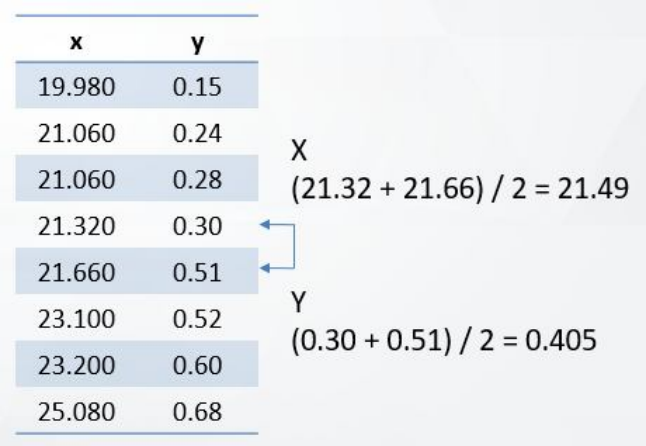

Figura 11 Obtenção da mediana do grupo k1

4.4.3 K-MEDOIDS

O método *k-medoids* tem a particularidade de que o centro do grupo sempre será um elemento do próprio grupo, sendo assim, sempre após a definição do centro do grupo, é feita a atualização desse centro para a posição do elemento do conjunto que estiver mais próximo a ele.

As **Figuras 12 e 13** ilustram a inicialização do centro de grupo e a atualização do centro de grupo após sua inicialização no passo 2, sucessivamente.

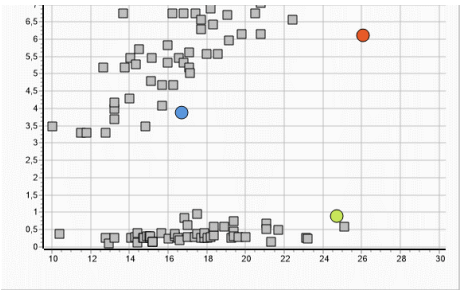

Figura 13 Inicialização dos centros de grupos

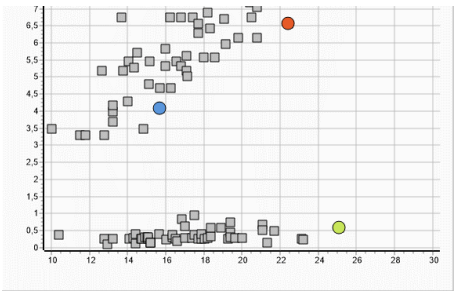

Figura 12 Centros de grupos após a atualização para o elemento mais próximo.

Sendo assim, após a obtenção do centro do grupo pela média, seguindo o método *k-means*, é feita a atualização para o elemento que estiver mais próximo.

As **Figuras 14 e 15** ilustram, o cálculo do centro de grupo k1 pela média e a atualização do centro de grupo para a posição do elemento que está mais próximo a ele, sucessivamente.

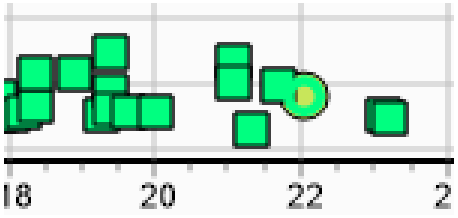

Figura 14 local do centro de grupo pela média

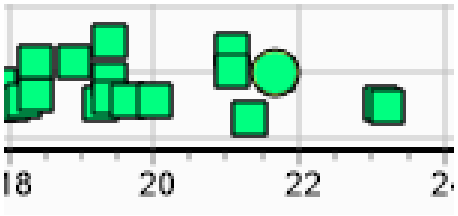

Figura 15 local do centro de grupo após atualização para o elemento mais próximo

4.5 REPETIÇÃO

Repetir os passos 3, 4 até que os centro de grupos *k* estejam fixados ou o número máximo de iterações seja atingido.

A **Figura 16** mostra o resultado da execução, com os grupos 3 separados.

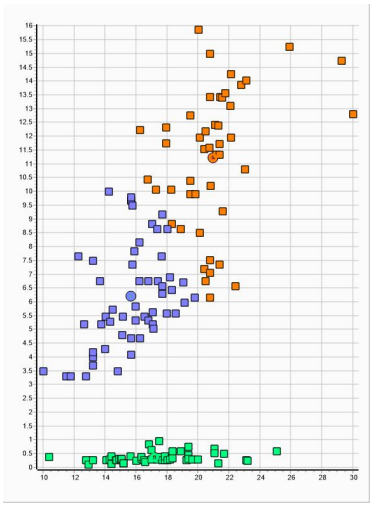

Figura 16 Resultado da execução do método

5 CONCLUSÃO

Este trabalho apresentou um detalhamento sobre 3 métodos de agrupamento de dados, *k-means*, *k-medians* e *k-medoids*, foram descritas as etapas para obtenção do agrupamento, suas medidas de dissimilaridade, tipo de dados permitidos e para cada um dos métodos foram descritos seus conceitos, vantagens e desvantagens.

Baseado no detalhamento dos métodos e nos resultados obtidos, é possível concluir que todos os métodos são eficientes na obtenção de agrupamentos, porém seu uso requer um bom conhecimento sobre a teoria do método aplicado, conjunto de dados estudado, tipo de dados e sobre a medida de dissimilaridade utilizada. A falta de conhecimento de qualquer um destes pontos levará a resultados distorcidos e sem precisão, como por exemplo a escolha do número em que o conjunto de dados será dividido, se escolhido de forma errada poderá gerar um grupo vazio ou com poucos elementos.

Verificou-se também que, embora os métodos para agrupamento de dados em geral, tenham passado por diversos avanços, ainda não existem métodos gerais que sejam adaptáveis a vários tipos de dados e geometrias dos conjuntos de dados, sendo assim há muita pesquisa para ser feita, principalmente para que o seu uso possa ser expandindo para outras áreas e para usuários com conhecimento mais superficial.

REFERÊNCIAS

BERRY, M. J. A., & LINOFF, G. **Mastering Data Mining: The Art and Science of Customer Relationship Management.** 1999.

FAYYAD, U. M., DJORGOVSKI, S. G., WEIR, N. **Automating the analysis and cataloging of sky surveys.** 1996.

KAUFMAN, L., ROUSSEEUW, P.J. **Finding Groups in Data: An Introduction to Cluster Analysis.** 1990.

MACQUEEN, J.B. **Some Methods for Classification and Analysis of Multivariate Observations**. 1967.

MAHDAVI M., ABOLHASSANI H. **Harmony K-means algorithm for document clustering**. 2009.

HAMERLY, G., ELKAN, C. **Learning the K in K-Means NIPS.** 2003.

KAUFMAN, L., ROUSSEEUW, P.J. **Clustering by means of Medoids, in Statistical Data Analysis Based on the L_1–Norm and Related Methods.** 1987.

BUBECK S., MEILA, M., VON LUXEMBOURG, U. **How the Initialization Affects the Stability of the K-Means Algorithm**. 2012.

JAIN, A.K., TOPCHY, A., LAW, M.H.C., BUHMANN, J.M. **Landscape of clustering algorithms.** 2004.

LICHMAN, M. UCI **Machine Learning Repository** [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2013.